MVA2005 IAPR Conference on Machine VIsion Applications, May 16-18, 2005 Tsukuba Science City, Japan

3-10

Robust Invariant Features for Object Recognition and Mobile Robot Navigation

Zhe Lin, Sungho Kim and In So Kweon
Dept. of EECS, Korea Advanced Institute of Science and Technology
373-1, Guseong-dong, Yuseong-gu, Daejeon 305701
{limcher, shkim}@rcv.kaist.ac.kr, iskweon@kaist.ac.kr

Abstract

In this paper, we present a novel local feature detector for the object recognition and robot navigation applications. The proposed algorithm extracts highly robust and repeatable features based on the key idea of tracking and grouping multi-scale interest points and selecting a unique representative structure with the strongest response in both spatial and scale domains. Weighted Zernike moments are used as the local descriptor for feature representation. The experimental results and performance evaluation show that our feature detector has high repeatability and invariance to large scale, viewpoint and illumination changes. The efficiency and usefulness of the proposed feature detection method are also confirmed by the excellent performance on object recognition and indoor topological navigation.

1 Introduction

Recently, the use of local features in the context of object recognition and vision-based robot localization and mapping has been successful due to their invariance and power to handle occlusions and background clutters.

The breakthrough work on local features [1] addresses the use of local invariants for content based image retrieval by combining the advantages of model-based and appearance-based approaches, and this has proved to be a very efficient approach. Later, many researches have been done based on this approach or its generalizations. Lindeberg has proposed an automatic scale selection mechanism which finds 3D maxima of Laplacian of Gaussian (LoG) filter in the scale space representation. Similarly, Lowe has presented an efficient feature detector – Scale Invariant Feature Transform (SIFT) [3] by searching for the 3D maxima of Difference of Gaussian (DoG) filter in the pyramid scale space. Due to its computational efficiency, SIFT has been used in many vision applications such as object recognition, indexing and mobile robot localization. It has been refined to reduce the noise sensitivity and edge effect in [4]. Mikolajczyk and Schmid have proposed a Harris corner based scale invariant interest point detector. The detector searches for the maximum response of Laplacian over scales to estimate the characteristic scale. They later have generalized this algorithm to the affine invariance by an iterative method [6]. The difference from earlier works [7, 8] is that this method simultaneously adapts the location and scale instead of using fixed feature locations. The main drawback of these iterative approaches is high computational complexity. Slightly different techniques have been proposed in [9, 10] by directly extracting invariant regions based on local intensity information. Parallelogram or elliptical regions in [9] and maximally stable extremal regions in [10] are typical examples of this technique. But, for arbitrary scenes, the regions detected by these approaches are very inconsistent due to the large image intensity variations and only few of them can be matched.

Although some of the above techniques have shown good results and wide applicability, the development of highly robust visual features is still a challenging problem. Our feature detector is aimed at detecting more robust and repeatable features. It first tracks interest points in the scale space to obtain structure-wise feature representation. Next, according to the information from each group of points, shape adapted local invariant regions are extracted. Finally, rotation invariant weighted Zernike moments [11] are calculated on the normalized local image patches for the local description.

In Section II~III, we describe in the details the RIF[1] detector and descriptor. Section IV introduces the application of the proposed feature detection algorithm to the object recognition and mobile robot navigation systems. Section V presents the experimental results and finally, in Section VI, we summarize the paper with conclusions and future works.

2 RIF Detector

2.1 Multi-scale interest points

Given an input image, first, we incrementally smooth it with Gaussian kernel to construct the multi-scale image representation. Next, from this multi-scale representation, the second moment matrix is calculated at every pixel location in each scale level image[2]. Then, the multi-scale interest points are localized at local peaks of the normalized Harris measure in the image domains.

Figure 1b shows an example result of the multi-scale interest point detection. Interestingly, we can note the following important facts: 1) Number of interest points decreases as scale increases and interest points exist in a range of scales, 2) Interest point location varies slightly over scales. The higher the scale level, the bigger the possible range of point locations, 3) Interest points can be locally structure-wise grouped so that each group represents a local structure. Additionally, we have verified from

[1] For convenience, we name the proposed region detector as Robust Invariant Feature (RIF) detector and the features as RIFs.

[2] For computational efficiency, we use the 4th order recursive implementation of Gaussian filtering [11] in calculating multi-scale representation and second moment matrices, which accelerates the overall feature detection process significantly.

many experimental tests that the locus of the interest points is very stable to the image rotation, scale, viewpoint and illumination changes. We define the evolution of the interest points in the scale space as the Local Corner Signature (LCS). With the scale increasing, interest points from different scale levels can be largely intersecting. It gives rise to a question that how to separate these points or how to assign them to each local structure. For solving this ambiguity, we propose the following tracking and grouping algorithm which automatically classifies interest points into each local structure or LCS.

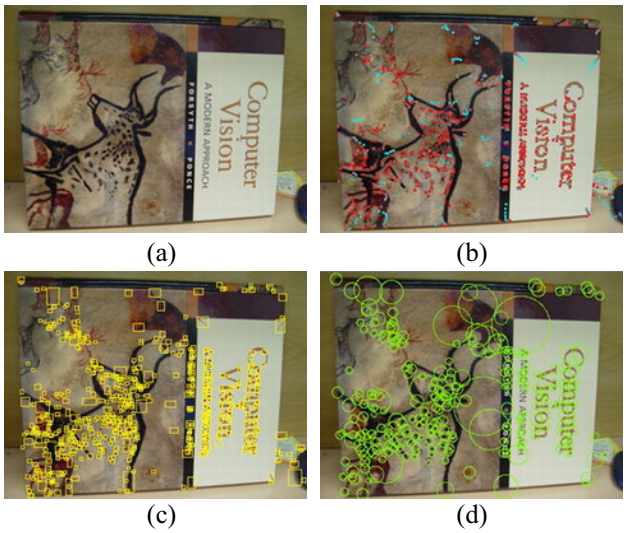
2.2 Tracking and grouping

Consider a set of multi-scale interest points extracted from the input image, $S = \{P_i^l = [x_i^l, y_i^l]^T \mid l = 1,...L, i = 1,...N_l\}$, where P_i^l denotes the i th interest point in the l th scale level, L denotes total number of scale levels and N_l denotes the number of interest points in l th scale level.

Our goal here is to track to group these interest points for the local structure-wise feature representation. The tracking and grouping is done by the following incremental linking algorithm (figure 2a):

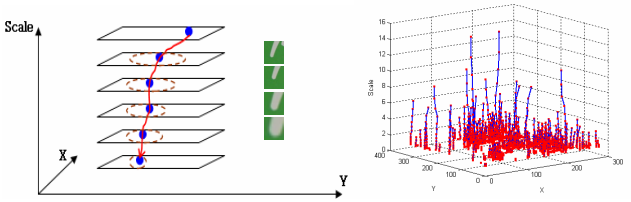
Pseudo code:

i) Initialize the current scale level as ($l \leftarrow L$) and current number of groups as ($k \leftarrow 0$).
ii) For the current scale level (l),
  a) Assign a new group G_k for every interest point without a link to the upper scale level and set ($k \leftarrow k+1$).
  b) For each interest point P_i^l in this level, search the nearest neighbor link in the lower level ($l-1$).
  If (the link exists within the range of uncertainty)
    Assign this link to the group corresponding to P_i^l.
  Else terminate the link at current level (l).
iii) Go to the next scale level ($l \leftarrow l-1$) and iterate ii ~ iii, until all interest points are assigned to a group or current level is the lowest one ($l = 0$).
iv) Output the feature groups $G_1, G_2, ...G_K$.



(a) (b) (c) (d)

Figure 1. Feature detection process (a) original image, (b) multi-scale interest point detection result (interest points detected from different scale levels are marked by different color pixels), (c) grouping result (each yellow box represents a group of features), (d) scale adaptation result (each green circle represents a locally adapted feature with the strongest response of normalized Harris measure over scales).



(a) (b)

Figure 2. Tracking and grouping (a) a sketch of interest points tracking, (b) LCSs (red square marks represent the interest points and blue curves represent the corner evolution routes.

Table 1. Experimental test for parameter k

k	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Grouping Error (%)	10.4	4.8	1.2	2.5	3.9	5.2	7.3	9.6

The idea is to cluster those multi-scale interest points corresponding to the same local structure. The linking process is gradually propagating from the highest scale level to lower scale levels based on the principle of nearest neighbor searching within the ranges of uncertainty. The tracking of interest points continues until no corresponding links within the allowed uncertainty ranges. The uncertainty range propagation is very important here, as it largely affects the grouping results. For example, when many textures existing in the image, the relative distances between local structures can be very small, hence many false groupings can be generated, which consequently decreases the number of correct estimations. We approximate the corner propagation rule as the following exponential model:

$$R(s,l) = k * s^l \quad (1)$$

where $R(s,l)$ is the radius of uncertainty regions, k is a constant factor, s is the parameter of the function and l is the scale level index. The parameter s is chosen as the same as the scale factor σ between consecutive scale levels. We have tested various values of k with respect to the resulting grouping error (table 1). The best value is obtained at $k = 0.5$. Figure 2b and 1c show the example result of grouping with $k = 0.5$.

2.3 Scale Adaptation

We use the normalized Harris measure for scale selection. This measure naturally fits the unified framework of searching for the strongest response in both spatial and scale domains. In addition, by the use of tracking and grouping algorithm, the ambiguity and inaccuracy in scale selection can be reduced maximally. We observe the trace of normalized Harris measure responses along the LCSs and search for the local peaks in the trace. Then, we select a unique representative scale at the strongest peak point so that the corresponding region appears the *most corner-like structure* (figure 1d). We can further estimate the exact scale by fitting parabola to each selected peak R_N.

$$s^* = \frac{s_{l+1} + s_{l-1}}{2} - \frac{\sigma(s_{l+1} - s_{l-1})(R_{l+1}^2 - R_{l-1}^2)}{R_{l+1}^2 - (\sigma^2 + 1)R_l^2 + R_{l-1}^2} \quad (2)$$

The initial feature points are re-localized based on the estimated scale s^* and the sub-pixel accuracy in localization is obtained by weighted average of contributions from eight neighboring points.

3 RIF Descriptor

For description, we first normalize the extracted patches to the canonical 10×10 circles. We assume the conventional gray-level linear illumination change model. Based on this simple model, we normalize the image patches by linearly shifting its mean and variance to fixed values. In this way, the scale and the offset can be eliminated to get the illumination normalized patch. For the description, weighted Zernike moments are used [11]. Zernike moments have superior properties in terms of image content representation, information redundancy and noise characteristics [14] so they can be reliably used in the recognition problem. It is defined over a set of complex polynomials which form a complete orthogonal set over the unit disk. Zernike moments are calculated by projecting the image intensity onto these orthogonal basis functions. Gaussian window is used to weight the image patch before calculating the descriptors. Moreover, since the individual components are uncorrelated, Euclidean distance can be used as the similarity measure for matching. The weighted Zernike moments are calculated as:

$$A_{nm} = \frac{n+1}{\pi} \sum_{x=0}^{N-1} \sum_{y=0}^{M-1} W(x,y) f(x,y) [V_{nm}(x,y)]^*$$

$$V_{nm}(x,y) = R_{nm}(x,y) \exp(jm \tan^{-1}(y/x)) \qquad (3)$$

$$R_{nm}(x,y) = \sum_{s=0}^{(n-|m|)/2} (-1)^s \frac{(n-s)!}{s! \left(\frac{n+|m|}{2} - s\right)! \left(\frac{n-|m|}{2} - s\right)!} (x^2 + y^2)^{(n-2s)/2}$$

where $V_{nm}(x,y)$ is orthogonal basis function, $R_{nm}(x,y)$ is radial polynomial, $f(x,y)$ is image function and $W(x,y)$ is the Gaussian weight. We use the fast implementation [15] to calculate Zernike moments.

4 Applications

The object recognition system consists of on-line and off-line parts. In the off-line learning stage, RIFs are detected from the model images and the resulting descriptors are stored in the database. In the on-line recognition stage, features detected from the input and model images are pushed into the searching engine to find the most similar match. The Euclidean distance is used to evaluate the similarity between descriptors. We use Approximate Nearest Neighbors (ANN) search algorithm and probabilistic voting technique for efficient DB indexing. Final verification stage ensures the recognition result to be correct by estimating the optimal homography and counting inliers and outliers. Consequently, the relative image image/object pose can be optimally estimated.

The key to the navigation is the estimated image/object pose information. We first manually drive the robot in its workspace and capture images at representative locations that the robot is expected to do critical motion. Then, the group of RIFs are extracted from these model images and stored as the scene landmarks. Finally, through our recognition and relative pose estimation system, the robot motion is planned to iteratively correct its motion and converges to the optimal pose matching with database pose. The iterative pose converging is based on the estimated landmark ID, verified optimal homography and estimated affine transformation. The robot motion is controlled by translation t and optimal affine transformation A^* as indicated above.

$$t = \begin{pmatrix} t_x \\ t_y \end{pmatrix} \rightarrow \begin{cases} \text{Rotation correction} & \|t\| > \tau \\ \text{Forward} & \|t\| < \tau \text{ and } \det(A^*) \; I) \; \varepsilon \\ \text{Stop} & \|t\| < \tau \text{ and } \det(A^*) \; I) \; \varepsilon \end{cases} \qquad (4)$$

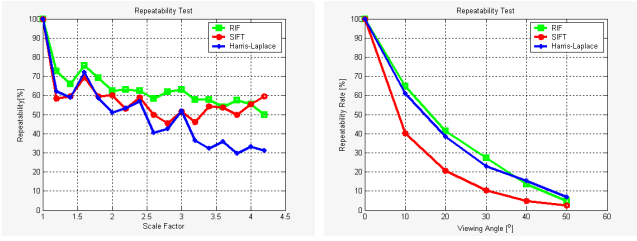
5 Experimental Results

5.1 Performance Evaluation of RIF Detector

We have tested the performance of the proposed feature detector under various image variations. The Hangul image set is selected as the experimental samples. Figure 3 shows some of the detection results. It shows that the RIFs are very consistent to large scale and illumination changes. For the viewpoint variation sequence, features are robust in a range of viewing angles ($-40^\circ \quad \theta \quad 40$).

We use the repeatability criterion [13] to evaluate the performance of the RIF detector. The sequences used in the experiment are selected from INRIA image database. Figure 4a shows that our proposed detector obtained higher repeatability rates than SIFT and Harris-Laplace (H-L) detector over the scale range of 1 to 4. The evaluation under viewpoint change shows that our detector has a slightly better repeatability rates than Harris-Laplace detector (figure 4b). Although our detector shows a high repeatability and robustness, it is still very sensitive to the large viewing angle changes. Our current work is focusing on the affine generalization in order to cover the limitation to large viewpoint changes.

Figure 3. RIF detection results under scale, viewpoint and illumination changes.



(a) (b)

Figure 4. Repeatability comparison of the proposed detector to SIFT and H-L detector under (a) scale changes, (b) viewpoint changes.

5.2 Results for recognition and robot navigation

We have tested the recognition performance for 483 images of 20 different objects (figure 5a) from KAIST Recognition DB. Figure 5b and 5c show examples for recognition under large image variations. The quantitative performance is evaluated for all the database images and the recognition rate (%) is compared with other two feature-based approaches (table 2). As the result, our detector shows the best recognition performance under scale changes, illumination changes and occlusions. But, for the viewpoint change sequences, our detector shows slightly lower recognition rate 95.8%. This is because the features are currently adapting shapes only to the uniform scales.

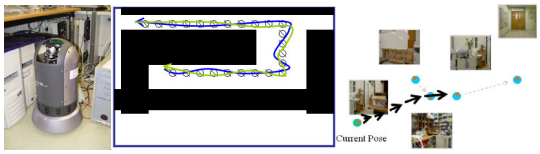
We have used KASIRI-IV robot as the experimental platform (figure 6a) which is a wireless commanded network based robot with single USB camera. Our experiment is performed in the indoor lab environment. Figure 6b, 6c shows the robot navigation path, environment and the topological node distribution. Our current navigation algorithm uses 16 different images as the landmark images. Some of those images are shown in figure 7a. Figure 7b shows the scene recognition and pose estimation results. We can see the system correctly recognizes landmarks even in the cases of the severe occlusion and scale change. In navigation experiment, the robot correctly achieved its goal to the final destination with 90% success rate among over 20 trials.

(a) (b) (c)

Figure 5. Recognition results (a) model images, (b) for large scale change, (c) for large viewpoint change.

Table 2. Experimental test for parameter k .

Recognition Rate (%)	Scale changes	Viewpoint changes	Illumination changes	Occlusions
RIF	63/63 (100%)	115/120 (95.8%)	80/80 (100%)	60/60 (100%)
Multi-scale Harris	61/63 (96.8%)	116/120 (96.7%)	80/80 (100%)	60/60 (100%)
SIFT	63/63 (100%)	100/120 (83.3%)	76/80 (95%)	56/60 (93.3%)



(a) (b) (c)

Figure 6. Experimental setting (a) KASIRI-IV Robot, (b) navigation environment and path, (c) topological nodes.

(a) (b)

Figure 7. Navigation results (a) landmarks, (b) examples of landmark recognition and pose estimation.

6 Conclusion and Future Works

We have developed a new feature detector and shown its application to object recognition and mobile robot navigation. Our detector, from the scale space interest point propagation to tracking and scale adaptation, basically models the fundamental knowledge on visual corner perception. Various experimental results have shown that our detector can generate highly robust and reliable features which can be efficiently used in the recognition. Our further application to the mobile robot navigation also proved its practical usefulness. Although our proposed feature detector showed a good performance, it can not handle well large viewpoint changes due to the large distortion of scale invariant circular regions. We are currently generalizing the proposed detector to the affine invariant feature which can handle large viewpoint changes.

Acknowledgement

This research is partially supported by Ministry of Information and Communications (MIC) and NRL (Code# M1-0302-00-0064) of MOST, Korea.

References

[1] C. Schmid and R. Mohr. “Local grey-value invariants for image retrieval,” *IEEE Transactions on PAMI*, 19(5): 530–534, May 1997.

[2] T. Lindeberg. “Feature detection with automatic scale selection,” *IJCV*, 30(2):79–116, 1998.

[3] D. G. Lowe. “Object recognition from local scale-invariant features,” *In Proc. of the 7th ICCV, Kerkyra, Greece*, pages 1150–1157, 1999.

[4] D. G. Lowe. “Distinctive image features from scale invariant keypoints,” *IJCV*, 60(2):91-110, 2004.

[5] K. Mikolajczyk and C. Schmid. “Indexing based on scale invariant interest points,” *In Proc. of the 8th ICCV, Vancouver, Canada*, pages 525–531, 2001.

[6] K. Mikolajczyk and C. Schmid. “An affine invariant interest point detector,” *In Proc. of the 7th ECCV, Copenhagen, Denmark*, volume I, pages 128–142, May 2002.

[7] T. Lindeberg and J. Garding. “Shape-adapted smoothing in estimation of 3-D shape cues from affine deformations of local 2-D brightness structure,” *Image and Vision Computing*, 15(6):415– 434, 1997.

[8] A. Baumberg. “Reliable feature matching across widely separated views,” *In Proc. of the International Conference on CVPR, South Carolina, USA*, pages 774–781, 2000.

[9] T. Tuytelaars and L. Van Gool. “Wide baseline stereo matching based on local, affinely invariant regions,” *In The 11th BMVC, U. of Bristol, UK*, pages 412–425, 2000.

[10] J. Matas, O. Chum, M. Urban and T. Pajdla. “Robust wide baseline stereo from maximally stable extremal regions,” *In The 13th BMVC, Cardiff U., UK*, pp. 384–393, 2002.

[11] S.H. Kim, I.C. Kim and I.S. Kweon. “Probabilistic model-based object recognition using local zernike moments,” *In IAPR Workshop on Machine Vision Applications, Nara, Japan*, Dec. 11-14, 2002.

[12] R. Deriche and G. Giraudon. “A computational approach for corner and vertex detection,” *IJCV*, 10(2):101–124, 1993.

[13] C. Schmid, R. Mohr, and C. Bauckhage. “Evaluation of interest point detectors,” *IJCV*, 37(2):151–172, 2000.

[14] T. Drummond and R. Cipolla. “Real-time visual tracking of complex structures,” *IEEE Transactions on PAMI*, 24(7): 932 –946, July 2002.

[15] Chee-Way Chong, P. Raveendran, R. Mukundan. “A comparative analysis of algorithms for fast computation of Zernike moments,” *Pattern Recognition* 36(3): 731-742, 2003.